# IMPORTANCE SAMPLING THROUGH DISTANCE MEASURES FOR ESTIMATING THE PROBABILITY OF A RARE EVENT.

A thesis submitted

in partial fulfilment of the requirements

for the Degree of

MASTER OF TECHNOLOGY

by

P. GOPATHY

to the

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

JANUARY 1991.

## ACKNOWLEDGEMENT

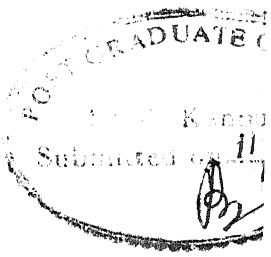
I am immensely indebted to Dr. Rakesh Kumar Bansal for having taught me the rigorous foundations of probability theory. I am also extremely thankful to him for his valuable guidance throughout the course of this thesis work.

Friends Prabhu Manyem, Kalyan Raman and Haja Mohideen helped in giving shape to this thesis and to them I remain indebted. Thanks to all my friends who kept me good company during my M.Tech.

My thanks are also due to Mr. L.S. Bajpai for his efficient typing and kind cooperation.

P. Gopathi

EE-1991-M-GOP-IMP



# CERTIFICATE

It is certified that the work contained in the th
entitled IMPORTANCE SAMPLING THROUGH DISTANCE MEASURES
ESTIMATING THE PROBABILITY OF A RARE EVENT by P. Gopathy,
been carried out under my supervision and that this work
not been submitted elsewhere for a degree.

January, 1991.

Dr. R. K. Bansal
Department of Electrical Engineeri
I.I.T. Kanpur

# ABSTRACT

The problem of reducing the sample size in the Monte-Carlo estimation of the probability of a rare event through importance sampling, a variance-reduction technique, has an optimal solution that is degenerate. Constrained optimal solutions have, therefore, been obtained through ad hoc approaches in many specific contexts.

In this thesis, guided by Kobayashi's theorem on the simultaneous minimization of all Ali-Silvey distances by the so called least favourable pair (in terms of Bayes's risk) in a composite binary hypothesis testing problem, constrained optimal solutions that minimise the variance of the importance sampling estimator are given for some group and exponential families. Using a lemma of Huber (1965) for a given pair to be least favourable, the asymptotic optimality of the biasing distribution obtained by a shift through threshold is established for the location family.

In the second part, Importance Sampling for Hall's Minimum Probability Ratio Tests (MPRTs) is studied in the spirit of Siegmund's results for Sequential tests.

# CONTENTS


| | | PAGE NO. |
|---|---|---|
| CHAPTER I | INTRODUCTION | 1 |
| CHAPTER II | PROBLEM FORMULATION | 7 |
| | 2.1 Importance Sampling Distance | 7 |
| | 2.2 Kobayashi's Theorem | 12 |
| | 2.3 Problem Statement | 15 |
| CHAPTER III | LOCATION, SCALE, LOCATION-SCALE AND SINGLE PARAMETER EXPONENTIAL FAMILIES | 21 |
| | 3.1 Minima of Kullback-Leibler Distance | 21 |
| | 3.2 Performance Comparison | 31 |
| | 3.3 Asymptotic Optimality of Shift Through Threshold | 35 |
| CHAPTER IV | SEQUENTIAL TESTS | 41 |
| | 4.1 Hall's Minimum Probability Ratio Tests (MPRTs) | 41 |
| | 4.2 Importance Sampling for MPRTs | 44 |
| CHAPTER V | CONCLUSION | 53 |
| | REFERENCES | 55 |
| | APPENDIX-A | 58 |

# CHAPTER - I

## I. INTRODUCTION

The Monte-Carlo estimation of the probability of a rare event is an often used technique in recent years, to characterize system performance in many contexts. Though only very few of these situations may claim full justification for the use of the Monte-Carlo method and are the ones where the complexity of the problem discourages analytical solutions because of operations involving many interacting random factors and where the analytical intractability of such problems may suggest the use of simulation [22, 23], still, the class of problems where Monte-Carlo methods are justifiably used is rather large. In communications and signal processing theory, the estimation of a false alarm rate [11], probability of a rare event in a stochastic algorithm[3], expectations of functionals of Markov chains, especially the probability of overflow in queuing networks etc., the estimation of bit error rate in a communication system [2, 12-14, 16, 18] and the estimation of error probabilities for sequential tests [19, 20] are examples of such problems. A typical case of bit error estimation for a communication system serves to exemplify the analytical intractability of such problems. Direct evaluation of bit

error rates for non linear systems perturbed by non-Gaussian noise is generally very difficult since probability densities on the output space may not be easily obtainable in closed forms. In such situations, one resorts to developing either certain bounds on the errors or certain assumptions that may allow the use of a central limit theorem. In the former case, the tightness of the bounds and in the latter case, the percentage error incurred are not always known and in some cases, have been known to grow exponentially with the number of observations [Orsak and Aazhang, 1990]. Hence the use of Monte-Carlo estimation techniques may be justified in such contexts.

In all such situations, the quantity to be evaluated, usually an error probability, is extremely small, of the order of $10^{-6}$ or less. Given a confidence interval of say, 95%, as shown in the next chapter, for an error probability $p_e$ one requires approximately $10/p_e$ simulations. Hence, for $p_e < 10^{-6}$, an inordinately large number of simulations, of the order of $10^7$ are required to attain the given accuracy. Apart from taxing the system used for simulation, this number can very well exceed the period of the random number generator used. Importance Sampling is a variance reduction technique that

aims at concentrating the simulation effort towards drawing the majority of samples from that region in the observation space which contributes most to the true value of the estimate. The principle involved here is to distort or modify the input random process in order to make the original low probability events, which influence the value of the estimate most, to occur more frequently. This action is then compensated by weighting the events appropriately and thus removing the bias introduced earlier. Kahn (1953) first studied this technique in a general framework. In communication and signal processing theory, this technique was first introduced by Shanmugam and Balaban (1980) and has been widely studied since [2, 3, 4, 11-18, 23] in different contexts.

Since Importance Sampling principally aims at drawing more samples from the"Important" region, the original input probability density is substituted with a biasing density which assigns more weight to this important region and samples are now drawn from this density. Ideally, one would like to put zero mass over the complement ("unimportant") region and normalise the given density. But the normalising constant will then be exactly equal to the quantity to be evaluated! This intuitive picture tells us that the optimal biasing density

what is catch-22?



will be degenerate; it puts the problem of identifying the best biasing density in a typical catch-22 situation.

In all the previous works, some parameteric family appropriately chosen for the specific problem at hand was used as the constraint family of interest from which the best biasing member was indentified by minimising the upper bound on the variance of the estimator. Other ad hoc approaches like proving that a certain choice of the biasing density gives a large reduction in the variance have also been employed. In Orsak and Aazhang (1990), a slightly different approach has been used. By defining Huber's mixture neighbourhoods around the optimal density and an arbitrarily chosen nominal density, of radii zero and some $\varepsilon > 0$ respectively, the member of the mixture neighbourhood around the nominal density which forms the least favourable pair with the optimal density is chosen as the sub-optimal biasing density. This choice is justified by Kobayashi's theorem [Kobayashi, 1967] that the least favourable pair in terms of Bayes's risk minimises all Ali-Silvey distance measures and hence also the so called Importance Sampling distance which is proportional to the variance of the Importance Sampling estimator. But since the least favourable densities are only truncated - normalised combinations of the

nominal densities, this solution is dependent on the degenerate optimal solution and is therefore unimplementable.

In this thesis, we apply Kobayashi's theorem from a different point of view. In Chapter II, we prove the non-existence of the least favourable pair, when one of the two neighbourhoods is shrunk to the optimal density and the other is a parametric family, through counter-examples. In Chapter III, with suitablejustification, we use the Kullback-Leibler distance measure instead of the Importance Sampling distance and come up with implementable solutions for some group families like the location, the scale, the location-scale families etc. and also the single-parameterexponential family. Further, all previous works have shown, either through brute-force simulation or by other empirical methods that the best biasing member within the shift family is the one with a shift that isequal to the threshold itself.Using Huber's (1965) lemma for a given pair to be least favourable, we have rigorously established the asymptotic optimality of the shift through threshold.

In Chapter IV, we consider Halls' Minimum Probability Ratio Tests, which is a very general class of sequential tests, including Wald's SPRT, t-tests, robust tests etc. as special cases, for the application of Importance Sampling. Siegmund's results for sequential test for a pair of simple hypothesis

indicate the use of the alternative as the biasing density for simulation under the null hypothesis as sub-optimal [Siegmund, 1976]. We therefore, test the same for Hall's MPRTs and show that the use of the alternative gives infinite reduction of the sample-size, asymptotically, over thecrude estimator. In the last chapter, we indicate the possible lines along which this result can be generalised for the multiple hypothesis case. Some improvements and modifications of the results by Orsak and Aazhang [1990] through the choice of some topological neighbourhoods e.g. total variational neighbourhood are also discussed.

# CHAPTER – II

## II. PROBLEM FORMULATION

In this chapter, we introduce the concept of importance sampling and outline the problem of identifying the best biasing member from a given class.

### 2.1 IMPORTANCE SAMPLING DISTANCE :

Let g be a real valued, Borel measurable, non decreasing function defined on real line. Let X be the 'input' random variable and Y = g(X) the 'output' random variable and the probability of a tail event of interest be expressed as

$$p_e = \int_{-\infty}^{\infty} I_1(y)\, p_Y(y)\, dy$$

where $I_1(y)$ is the indicator function of the event $[T_1, \infty)$. We can rewrite the expression for the error probability in terms of $P_X(x)$ as follows :

$$p_e = \int_{-\infty}^{\infty} I_1(g(x))\, p_X(x)\, dx = \int_{-\infty}^{\infty} I(x)\, p_X(x)\, dx \qquad (2.1)$$

where I(x) now is the indicator function of another tail set $[T, \infty)$. Here we have made use of the monotonicity of g(.).

Suppose $P_e$ in (2.1) is to be estimated through a crude Monte-Carlo method by the sample mean

$$\hat{p}_e = \frac{1}{N} \sum_{i=1}^{N} I(X_i)$$

where $X_i$ are iid random variables with density $p_X(x)$. The variance of this estimator is

$$\mathrm{Var}(\hat{p}_e) = \frac{1}{N} p_e(1 - p_e) \qquad (2.2)$$

If $\hat{p}_e$ is specified to lie within 10% of $p_e$ with a probability of say, 0.95, then the number of simulations required to achieve this may be computed using Chebyshev's inequality :

$$P\left\{|\hat{p}_e - p_e| \geq \frac{p_e}{10}\right\} \leq \frac{\mathrm{Var}(p_e)}{p_e^2/100} = \frac{100}{p_e} \cdot \frac{(1-p_e)}{N}$$

Approximating $(1-p_e)$ by 1 and the probability of the confidence interval also by 1, we see that $N \geq \frac{100}{p_e}$ is required to claim this level of accuracy. If $p_e$ is of the order of $10^{-6}$ or less, then $N \geq 10^8$, which will both tax the system exorbitantly and may well exceed the period of the random number generator.

Kahn [1953] suggests the following alternative : Rewriting (2.1) as

What is absolutely continuous mean?



$$p_e = \int_{-\infty}^{\infty} I(x) \cdot \frac{p_X(x)}{p_X^*(x)} \cdot dP_X^*(x) \qquad (2.3)$$

where $P_X^*$ is a probability measure such that $P_X$ is absolutely continuous wrt it and $p_X^*$ is its density. We now call $p_X^*$ the "biasing" density and form the following estimate for $p_e$ :

$$\hat{p}_e^* = \frac{1}{N} \sum_{i=1}^{N} I(X_i) \frac{p_X(X_i)}{p_X^*(X_i)}$$

where $X_i \sim P_X^* >> P_X$. (Hereafter, we uniformly drop the suffix X). The variance of this new estimate is

$$Var(\hat{p}_e^*) = \frac{\int_{-\infty}^{\infty} I(x) \left[\frac{p(x)}{p^*(x)}\right]^2 dP^*(x) - p_e^2}{N} \qquad (2.4)$$

If, by a proper choice of $p^*(x)$, we can reduce the above variance such that

$$Var(\hat{p}_e^*) < Var(\hat{p}_e),$$

then we would have achieved a corresponding reduction in the number simulations.

**Theorem 1**

[Orsak and Aazhang, 90] : The choice of $p^*(x) = \frac{I(x)\, p(x)}{p_e}$ achieves the minimum variance for the importance sampling estimator. Let us call this the 'Optimal biasing density', $p_{opt}$.

Proof : Applying Jensen's inequality to the second moment in (2.4), we have

$$E_{p^*}\left[I(X) \cdot \frac{p(X)}{p^*(X)}\right]^2 \geq E^2_{p^*}\left[I(X)\frac{p(X)}{p^*(X)}\right]$$

$$= E^2_p\,(I(X)) = p_e^2$$

The equality is achieved iff $I(X) \cdot \frac{p(X)}{p^*(X)} = k$, a.s., where k is a constant. Integrating either side over $(-\infty, \infty)$, we have $k = p_e$. QED

This degenerate optimal solution can intuitively be explained as follows : We will achieve maximum reduction in sample-size if we take samples only from the "important" region i.e., $[T, \infty)$ and no samples at all from $(-\infty, T]$. Then, putting zero mass over this region or equivalently, truncating p(x) with an indicator and normalizing it, we get the optimal biasing density.

In all previous works, implementable solutions have been derived by choosing some constraint class $\mathbb{P}$ specific to the problem at hand and solving for

$$p_1^* = \arg \min_{p^* \in \mathbb{P}} \text{Var}\ (\hat{p}_e^*).$$

In Orsak and Aazhang [1990], $\mathbb{P}$ has been chosen as the mixture neighbourhood.

Let us now rewrite the second moment in (2.4) as follows :

$$\int_{-\infty}^{\infty} I(x) \left[ \frac{p(x)}{p^*(x)} \right]^2 dP^*(x) = p_e^2 \int_{-\infty}^{\infty} \left[ \frac{I(x)\, p(x)}{p_e \,.\, p^*(x)} \right]^2 dP^*(x)$$

$$= p_e^2 \int \left[ \frac{p_{opt}(x)}{p^*(x)} \right]^2 dP^*(x) \qquad (2.5)$$

Let us define a distance measure $d(P_1, P_2)$ between the two probability measures $P_1$ and $P_2$ as follows :

$$d(P_1, P_2) \overset{\Delta}{=} \int_{-\infty}^{\infty} \left( \frac{p_1}{p_2} \right)^2 dP_2, \quad P_1 << P_2 \,. \qquad (2.6)$$

Then, $d(P_1, P_2)$ is an Ali-Silvey distance measure [Ali and Silvey, 1966]. See Appendix A. Using (2.5) and (2.6) one can write (2.4) as

$$Var(\hat{p}_e^*) = \frac{p_e^2 \,(d(P_{opt}, P^*) - 1)}{N}$$

Let us call $d(P_{opt} P) \overset{\Delta}{=} d_{IS}(P_{opt}, P)$, the "Importance Sampling distance" as in Orsak and Aazhang,(1990). Using Jensen's inequality or the properties of Ali-Silvey distance measures, we can show that $0 \leq d_{IS} \leq 1$. Again, $d_{IS}(P_{opt}, P^*) = 1$ iff $P^* = P_{opt}$, and then the variance is zero, the minimum possible.

Thus, we reformulate our original problem as follows : solve for

$$P_1^* = \arg \min_{P^* \in \mathbb{P}} d_{IS}(P_{opt}, P^*), \qquad (2.7)$$

where $\mathbb{P}$ is the given contraint class.

## 2.2 KOBAYASHI'S THEOREM :

Kobayashi's theorem gives a relationship between distance measures and the robust detection problem. We use this relationship to further study the importance sampling problem.

Let $(\Omega, \mathbb{F})$ be a measurable space and $P_o$ and $P_1$, two probability measures on it, with densities $p_o$ and $p_1$ respectively, wrt some measure $\mu$. Unknown deviations from the nominal pair can be modelled by blowing up $P_i$ into composite hypotheses

$$\mathbb{P}_i = \{Q \mid Q = (1-\varepsilon_i) P_i + \varepsilon_i H_i, \; H_i \in \mathbb{H}\}, \; i = 0,1 \qquad (2.8)$$

where $0 \leq \varepsilon_i < 1$ are the "radii" of the two classes and $\mathbb{H}$ is the class of all probability measures on $(\Omega, \mathbb{F})$. Let $\phi(x)$ be the conditional probability of acceptance of $P_1$ given x. If a loss of $L_i > 0$ is incurred when $\mathbb{P}_i$ is falsely rejected, then the expected loss or risk under $Q_i'$ is

$$R(Q_i', \phi) = L_i \, |i - E_{Q_i'}(\phi)|, \quad i = 0,1$$

Where $Q_i'$ is the true underlying distribution. The classical minimax, Neyman-Pearson and Bayes's criteria are now given a universal minimax spirit :

i) minimise $\max_{i=0,1} \sup_{Q_i'} R(Q_i', \phi)$

ii) minimise $\sup_{Q_i'} R(Q_i', \phi)$, subject to $\sup_{Q_o'} R(Q_o', \phi) \leq \alpha$

iii) minimise $\sup_{Q_1', Q_o'} \{ \lambda_o R(Q_o', \phi) + \lambda_1 R(Q_1', \phi) \}$

where $\lambda_i$'s are the priors.

For the test of $P_o$ against $P_1$ given by

$$\phi(X) = \begin{cases} 1, & \dfrac{p_1(x)}{p_o(x)} > \gamma \\ \xi, & \dfrac{p_1(x)}{p_o(x)} = \gamma \\ 0, & \dfrac{p_1(x)}{p_o(x)} < \gamma \end{cases}$$

Huber (1965) has shown that there exists a pair of distributions $Q_i \in \mathbb{P}_i$ such that

$$R(Q_i', \phi) \leq R(Q_i, \phi)$$

and hence the minimax, Neyman-Pearson and the Bayes's criteria for the test of $Q_o$ and $Q_1$ are exactly equivalent to conditions (i), (ii) and (iii) above, respectively. This "least favourable pair" was also given by Huber :

$$q_o(x) = \begin{cases} (1-\varepsilon_o)\, p_o(x), & p_1(x)/p_o(x) < C'' \\ (1/C'')\,(1-\varepsilon_o)\, p_1(x), & p_1(x)/p_o(x) \geq C'' \end{cases} \qquad (2.8)$$

$$q_1(x) = \begin{cases} (1-\varepsilon_1)\, p_1(x), & p_1(x)/p_o(x) < C' \\ C'\,(1-\varepsilon_1)\, p_o(x), & p_1(x)/p_o(x) \geq C' \end{cases}$$

where $o \leq C' < C'' \leq \infty$.

We state another useful result here as a Theorem.

**Theorem 2** [See Huber (1965) Lemma 2 and the discusion that follows] : If for any $Q_i' \in \mathbb{P}_i$, $i = 0,1$, and any real $t$, we have

$$Q_o'[q_1/q_o < t] \geq Q_o[q_1/q_o < t] \geq Q_1[q_1/q_o < t] \geq Q'[q_1/q_o < t],$$

then $(Q_o, Q_1)$ is the least favourable pair.

Note : $(Q_o, Q_1)$ is referred to as "least favourable in Bayes's sense" if

$$(Q_o, Q_1) = \arg \sup_{Q_o', Q_1'} \{\lambda_o R(Q_o', \phi) + \lambda_1 R(Q_1', \phi)\}$$

Theorem 3 $(Q_o, Q_1)$ is least favourable in Bayes's sense for all sets of priors $(\lambda_o, \lambda_1)$ if and only if

$$d(Q_o, Q_1) \leq d(Q_o', Q_1')$$

for all Ali-Silvey distance measures d.

For proof of this theorem, see Kobayashi (1967). Intuitively, a given pair will be least favourable only if they are "close" to each other and as far from the nominal measures as possible.

## 2.3 PROBLEM STATEMENT :

In Orsak and Aazhang (1990), the Huber neighbourhood (2.7) was defined as follows :

$$P_1 \overset{\Delta}{=} P_{opt}, \quad \varepsilon_1 \overset{\Delta}{=} 0, \quad \varepsilon_o > 0$$

and $P_o$, some arbitrary nominal density. Then, for some mixture or parametric (eg., shift class) class $\mathbb{P}_o$, the least favourable

density $q_o$ was written using (2.8). By Kobayashi's theorem (theorem 3, above), this minimises the Importance Sampling distance from $P_{opt}$, i.e.,

$$Q_o = \arg \min_{Q_o' \in \mathbb{P}_o} d_{IS}(P_{opt}, Q_o')$$

Hence $Q_o$ is an optimal solution for the Importance Sampling problem subject to the constraints used.

By a different use of Kobayashi's theorem, we now show that this is not so.

**Counter-Example 1 :** Consider the double-exponential Location family

$$f_{T'}(x) = \frac{1}{2} e^{-|x-T'|}, \quad T' \in (-\infty, \infty)$$

Let $f_o(x)$ be the density of the null hypothesis and the probability of false-alarm be

$$p_e = \int_T^{\infty} f_o(x)\, dx = e^{-T}.$$

This is the quanity to be evaluated by simulation (assumed unknown). Then,

$$p_{opt}(x) = \begin{cases} e^{T-x}, & x \geq T \\ 0, & \text{otherwise} \end{cases}$$

To choose the sub-optimal density from this family, let us compute the generic Importance Sampling distance and minimise it

$$d_{IS}\left(p_{opt}(x),\ f_{T'}(x)\right) = \begin{cases} 2\, e^{T-T'}, & T' \leq T \\ \frac{2}{3}\, e^{T'-T} + \frac{4}{3}\, e^{2(T-T')}, & T' \geq T \end{cases}$$

This is minimised by $T' = T$ over $T' \leq T$ and by $T' = T + \frac{1}{3}\ln 4$ for $T' > T$. Hence, the sub-optimal $T'$ is given by

$$T' = T + \frac{1}{3}\log 4 \qquad (2.9)$$

Let us also compute the Kullback-Leibler distance, which is also an Ali-Silvey distance, between $p_{opt}(x)$ and the generic member

$$d_{KL}\left(p_{opt},\ f_{T'}\right) = \int_{-\infty}^{\infty} \log\left[\frac{p_{opt}(x)}{f_{T'}(x)}\right] p_{opt}(x)\, dx$$

By direct calculation, we see that the $T'$ minimising the above is

$$T' = T + \log 2 \qquad (2.10)$$

Since the minima of these two Ali-Silvey distances occur at two different points, by Kobayashi's theorem, the least favourable density does not exist.

**Counter-Example 2 :** Again, for the single parameter exponential class

$$p_\theta(x) = \exp(\theta x - \Psi(\theta))\, p_o(x),$$

where $\Psi(\theta)$ is a convex function normalised such that $\Psi(o) = 0$, we can directly compute the importance sampling and the Kullback-Leibler distance and minimise them. Here

$$P_{opt}(x) = I_{[T,\infty)}(x)\, p_o(x)/p_e, \quad \text{where } p_e = \int_T^\infty p_o(x)\, dx$$

$d_{KL}$ is minimised by the $\theta_o$ that solves $\Psi'(\theta_o) = \mu_{1,opt}$.

where

$$\mu_{1,opt} = E_{p_{opt}}(X) \qquad (2.11)$$

and $d_{IS}$ is minimised by the $\theta_o$ that solves

$$\Psi'(\theta_o) = E_{p_{opt}}[X\, e^{-\theta_o X}]/E_{p_{opt}}[e^{-\theta_o X}] \qquad (2.12)$$

As we show later, these two are only asymptotically (i.e., for $T \longrightarrow \infty$) equal. Hence, for $p_e > 0$, the least favourable density does not exist.

We observe from the above two cases that though the minima occur at different parameter values for the two distance measures, for large T or small $p_e$, they are pretty close. This is obvious from 2.9, 2.10, in counter example 1 but for counter example 2 using 2.11 and 2.12, we do this by first obtaining

$$\lim_{T\longrightarrow\infty} \frac{\Psi'(\theta_o)}{T} = \lim_{T\longrightarrow\infty} \frac{\mu_{1,opt}}{T} = \lim_{T\longrightarrow\infty} \frac{\int_T^\infty x p_o(x)\, dx}{T} = 1$$

under some mild conditions on $p_o(x)$ (see Chapter 3) and then,

$$\lim_{T\longrightarrow\infty} \frac{E_{p_{opt}}[X\, e^{-\theta_o X}]/E_{p_{opt}}[e^{-\theta_o X}]}{T} = 1$$

Since the values of the parameters minimising these two distance measures are approximately equal for large T, in both the above cases, one can consider using the Kullback-Leibler distance measure instead of the Importance Sampling distance to find sub-optimal parameter estimates.

The above mentioned devious route of minimising the Kullback-Leibler distance and assuming that, asymptotically all other distance measures are minimised at the same point is further justified by a rigourous proof that we give in section 3.3 for the existence of the least favourable density for the shift class under the condition that $p_e \longrightarrow 0$.

The difficulty of working with the Importance Sampling distance or equivalently, the second moment of the Importance Sampling estimate is clearly illustrated by (2.12) above.

# CHAPTER - III

## LOCATION, SCALE, LOCATION-SCALE AND SINGLE PARAMETER EXPONENTIAL FAMILIES

In this chapter, we identify the best biasing members from the location (shift), scale, location (shift) - scale and single parameter exponential families. We also prove the asymptotic optimality for the first case.

### 3.1 MINIMA OF KULLBACK-LEIBLER DISTANCE :

The two counter examples in Chapter II show that atleast for the double-exponential location family and the single-parameter exponential family, for large T (small $p_e$), the Importance Sampling distance can be substituted with the Kullback-Leibler distance in equation (2.7). The main advantage resulting from this is the ease of handling. Further, as we shall see shortly, an implementable and convenient estimate of the parameter value that minimises the Kullback-Leibler distance can be obtained. If, for a given parametric family, the least favourable member for the corresponding $p_{opt}$ exists, then by Kobayashi (1967), we are fully justified in using this as the sub-optimal biasing density. Otherwise, with the counter examples in mind, we may assume that this is "close" to the sub-optimal density for large T and then cross check the result by actually computing asymptotically ($T \longrightarrow \infty$) the ratio

of the second moment for this Importance Sampling estimator to the second moment of the crude estimator.

In literature, so far, the following three families have been studied, under different contexts :

i) Shift : $p(x - T)$, $T \in (-\infty, \infty)$

ii) Scale : $\frac{1}{\sigma} p\left(\frac{x}{\sigma}\right)$, $\sigma \in (0, \infty)$

iii) Till : $\frac{e^{\lambda x} p(x)}{M(\lambda)}$, $\lambda \in (0, \infty)$

Motivated by this, we consider for the analysis outlined above, the location, scale, location-scale families and the single-parameter exponential family. First, we consider the example of the Gaussian family :

**Example :**

i) **Shift** : Let $p_o(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Then, $p_e = \int_T^\infty p_o(x)\, dx$ and

$$p_{opt}(x) = \frac{I_{[T,\infty)}(x)}{p_e \cdot \sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and $p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\theta)^2}{2}}$

$$d_{KL}\ (P_{opt},\ P_{\theta}) = \int_{-\infty}^{\infty} \log\left[\frac{p_{opt}(x)}{p_{\theta}(x)}\right] dP_{opt}(x)$$

This is minimised for $\theta_o = \mu_{1,opt} \overset{\Delta}{=} \int_{-\infty}^{\infty} x.p_{opt}(x)\,dx$. Since $p_e$ is unknown, $p_{\mu_{1,opt}}$ is also not known. However,

$$\lim_{T\to\infty} \frac{\mu_{1,opt}}{T} = \lim_{T\to\infty} \frac{\int_{T}^{\infty} x\, e^{-\frac{x^2}{2}}\,dx}{T\int_{T}^{\infty} e^{-\frac{x^2}{2}}\,dx} = 1. \qquad (3.1)$$

Alternatively, by using the well-known [Ref. no. 24, Feller (1966)] asymptotically tight bound on

$$Q(T) = \int_{T}^{\infty} e^{-\frac{x^2}{2}}\,dx \ :\ \left(\frac{1}{T} - \frac{1}{T^3}\right) e^{-\frac{T^2}{2}} \leq Q(T) \leq \frac{1}{T} e^{-\frac{T^2}{2}},$$

we can obtain the corresponding bound on $\mu_{1,opt}$ as

$$\mu_{1,opt} \in \left[T,\ \frac{T^2}{T-1}\right], \text{ for} \qquad (3.2)$$

$\therefore\ \mu_{1,opt} \sim T$ for large T. In other words, to estimate $p_e$, one uses $p_T(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-T)^2}{2}}$ as the biasing density. This

result is consistent with Orsak and Aazhang (1989), Wessel, Hall and Wise (1988) and Beaulieu (1990).

ii) **Scale** : Here, $p_o(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $p_{opt}(x) \frac{I_{[T,\infty)}(x)}{\sqrt{2\pi}.p_e} e^{-\frac{x^2}{2}}$

and $p_\sigma(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}}$

$$d_{KL}(P_{opt}, P_\sigma) = \log\left(\frac{\sigma}{k}\right) + \frac{\mu_{2,opt}}{2}\left[\frac{1}{\sigma^2} - 1\right]$$

and $\sigma_o^2 = \mu_{2,opt} = \int_{-\infty}^{\infty} x^2 p_{opt}(x)\, dx$

minimise the above Kullback-Leibler distance. Again,

$$\lim_{T\longrightarrow\infty} \frac{\sigma_o^2}{T^2} = 1.$$

Further the exact bounds on $\mu_{2,opt}$ are as follows :

$$\mu_{2opt} \in \left[1 + T^2 \ , \ 1 + \frac{T^4}{T^2 - 1}\right] \text{ for } T > 1 \qquad (3.3)$$

$$\therefore \quad \sigma_o^2 = \mu_{2,opt} \sim T^2 \qquad (3.4)$$

iii) **Shift and Scale** : $p_{opt}(x) = \frac{I_{[T,\infty)}(x)}{p_e} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

and $$p_{\theta,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

$$d_{KL}(P_{opt}, P_{\sigma,\theta}) = \log\left(\frac{\sigma}{k}\right) - \frac{\mu_{2,opt}}{2} + \frac{\mu_{2,opt}}{2\sigma^2} + \frac{\theta^2}{2\sigma^2} - \frac{\theta\mu_{1,opt}}{\sigma^2}.$$

$$\theta_o = \mu_{1,opt} \text{ and } \sigma_o^2 = \mu_{2,opt} - \mu_{1,opt}^2 \qquad (3.5)$$

minimise the above jointly.

Therefore, $\theta_o \sim T$ and $0 < \sigma_o^2 < 2$, for large T. This solution is modified in section 3.2 when the actual performance of these three estimators are computed.

In the following three sections, by imposing some conditions, we obtain estimates similar to the above for the general location and scale families. To that end, first we present the following lemma.

**Lemma 1**

For a density function f(X) satisfying

$$\int_T^\infty f(x)\,dx << T\ f(T) \text{ for large T,}$$

$$\mu_{1,opt} = \int_{-\infty}^{\infty} x\, p_{opt}(x)\,dx \sim T \text{ for large T, where}$$

$p_{opt}(x) = \frac{I_{[T,\infty)}(x)\ f(x)}{k}$ and $k = \int_T^\infty f(x)\,dx$ is the quantity

to be estimated.

**Proof :**

$$\lim_{T\to\infty} \frac{\mu_{1,opt}}{T} = \lim_{T\to\infty} \frac{\int_T^\infty x\ f(x)\,dx}{T\int_T^\infty f(x)\,dx}$$

$$= \lim_{\infty} \frac{-T\ f(T)}{\int_T^\infty f(x)\,dx - T\ f(T)} = 1$$

(by L'Hospital's rule and Leibnitz theorem).

**Remark :**

It is easily verified that the above condition is satisfied for the double exponential and Gaussian families.

**3.1.1 Shift family :**

Let $p_o(x) = f(x)$ and $p_\theta(x) = f(x-\theta)$.

$\therefore\ p_{opt}(x) = \frac{I_{[T,\infty)}(x)}{k} f(x)$ ; $k = \int_T^\infty f(x)\,dx$

is the quantity to be estimated. Let f(x) be such that

i) f(.) is differentiable and is normalised by f'(0) = 0

ii) $\varphi(x) = \dfrac{f'(x)}{f(x)}$ is convex and monotone decreasing over $[0,\infty)$.

Then $d_{KL}(P_{opt}, P_\theta)$ is minimised by the $\theta_o$ that solves

$$\frac{1}{k} \int_T^\infty \left\{ \frac{f'(x-\theta_o)}{f(x-\theta_o)} \right\} f(x)\, dx = 0$$

We solve for $\theta_o$ as follows :

$$\frac{1}{k} \int_T^\infty \left\{ \frac{f'(x-\theta_o)}{f(x-\theta_o)} \right\} f(x)\, dx = \frac{1}{k} \int_T^\infty \varphi(x-\theta_o)\, f(x)\, dx$$

$$\geq \varphi \left\{ \frac{1}{k} \int_T^\infty (x-\theta_o)\, f(x)\, dx \right\}$$

$$= \varphi\, (\mu_{1,opt} - \theta_o) = 0 \qquad (3.6)$$

when $\theta_o = \mu_{1,opt}$ Use has been made of Jensen's inequality and conditions (i) and (ii) above. Again,

$$\frac{1}{k} \int_T^\infty \left\{ \frac{f'(x-\theta_o)}{f(x-\theta_o)} \right\} f(x)\, dx \leq \frac{f'(T-\theta_o)}{f(T-\theta_o)} = 0 \qquad (3.7)$$

when $\theta_o = T$, by (i) and (ii) above.

But $\mu_{1,opt} \sim T$, for large T. (by lemma 1) (3.8)

Therefore, from 3.6, 3.7 and 3.8, we see that the Kullback- Leibler distance for this case is minimised by $\theta_o \sim T$, for large T.

Condition (i) and (ii) are satisfied for the generalised Gaussian family

$$f(X) = \frac{p}{2\Gamma(1/p)\, A(p)} \cdot \exp\left\{ - \left[ \frac{|X|}{A(p)} \right]^p \right\} \quad (3.9)$$

where

$$A(p) \overset{\Delta}{=} [\sigma^2\, \Gamma(1/p)/\, \Gamma(3/p)]^{1/2}, \text{ for all } p > 1.$$

See Orsak and Aazhang (1989).

### 3.1.2 Scale family :

Let $p_o(x) = f(x)$ and $p_\sigma(x) = \frac{1}{\sigma} f(\frac{x}{\sigma})$.

$$\therefore\ p_{opt}(x) = \frac{I_{[T,\infty)}(x)\, f(x)}{k}, \text{ where } k = \int_T^\infty f(x)\, dx$$

is the quantity to be evaluated.

The Kullback-Leibler distance $d_{KL}(P_{opt}, P_{\sigma})$ is minimised by the $\sigma_o$ which satisfies

$$\frac{1}{k} \int_T \left[ \frac{x}{\sigma_o} \right] \frac{f'(\frac{x}{\sigma_o})}{f(\frac{x}{\sigma_o})} f(x)\, dx = -1$$

We impose similar conditions on f here, as in the previous case :

i) f(.) is differentiable and is normalised such that f'(0) = 0

ii) $\varphi(y) = y \frac{f'(y)}{f(y)}$ is convex and monote decreasing over $[0,\infty)$.

Then,

$$\frac{1}{k} \int_T \left[ \frac{x}{\sigma_o} \right] \cdot \frac{f'(\frac{x}{\sigma_o})}{f(\frac{x}{\sigma_o})} f(x)\, dx = \frac{1}{k} \int_T \varphi(\frac{x}{\sigma_o}) f(x)\, dx$$

$$\geq \varphi \left\{ \frac{\mu_{1opt}}{\sigma_o} \right\} \quad \text{(by Jensen's inequality)}$$

If we set the r.h.s. above equal to - 1, then

$$\frac{1}{k} \int_T \left[ \frac{x}{\sigma_o} \right] \frac{f'(\frac{x}{\sigma_o})}{f(\frac{x}{\sigma_o})} f(x)\, dx \geq -1.$$

But for large T in view of lemma 1,

$$\phi\left\{\frac{\mu_{1,opt}}{\sigma_o}\right\} = \frac{T}{\sigma_o} \cdot \frac{f'(\frac{T}{\sigma_o})}{f(\frac{T}{\sigma_o})} = -1. \qquad (3.10)$$

Again,

$$\frac{1}{k}\int_T^\infty \left[\frac{x}{\sigma_o}\right] \frac{f'(\frac{x}{\sigma_o})}{f(\frac{x}{\sigma_o})} f(x)\,dx \le \frac{T}{\sigma_o}\frac{f'(\frac{T}{\sigma_o})}{f(\frac{T}{\sigma_o})} \qquad (3.11)$$

From 3.10 and 3.11 above, we conclude that $\sigma_o$ solving

$$\frac{T}{\sigma_o} \cdot \frac{f'(\frac{T}{\sigma_o})}{f(\frac{T}{\sigma_o})} = -1$$

asymptotically minimises the Kullback-Leibler distance in this case.

### 3.1.3 Single-parameter exponential family :

Referring to counter example 2 in Chapter II, we see that to estimate

$$k = \int_T^\infty p_o(x)\,dx,$$

$\mu_{1,opt} \sim T$, for large T.

The conditions in sections 3.1.1 and 3.1.2 are satisfied by the generalised Gaussian family in (3.9), for all $p > 1$. A condition equivalent to 3.1.1 (ii), namely that $-\log(f(x))$ is convex, has been imposed in Orsak and Aazhang (1989), but only a much weaker result, that is, $\theta_o = T$ proves better than the crude estimator (i.e. $\theta_o = 0$) has been obtained (through an ad hoc approach).

The results obtained for the Gaussian example initially are seen to be in agreement with those of sections 3.1.1, 3.1.2 and 3.1.3.

## 3.2 PERRFORMANCE COMPARISON:

In this section, we study the relative performance of the different biasing schemes outlined above, namely, shifting, scaling, shifting and scaling and also of the crude estimation method.

**Gaussian Example :**

Here, the second moment of the crude estimator is

$$\mu_{2,MC} = p_e = \frac{1}{\sqrt{2\pi}} \int_T^\infty e^{-\frac{x^2}{2}} dx \sim \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{T^2}{2}}}{T}, \quad (3.12)$$

Here we have used the asymptotically tight bound on Q(T) as in the previous section.

With $p_{\theta_o}(x)$, $\theta_o = \mu_{1,opt} \sim T$ as the biasing density,

$$\mu_{2,(shift=T)} = \int_T^\infty \left[ \frac{p_o(x)}{p_T(x)} \right]^2 p_T(x)\, dx$$

$$\sim \frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-T^2}}{2T} \quad (3.13)$$

With $p_{\sigma_o}(x)$, $\sigma_o^2 \sim T^2$ as the bising density

$$\mu_{2,(scale=T)} = \int_T^\infty \left[ \frac{p_o(x)}{p_{\sigma_o}(x)} \right]^2 p_{\sigma_o=T}(x)\, dx$$

$$\sim \frac{1}{\sqrt{2\pi}} \left[ \frac{T^2}{2T^2-1} \right] \cdot e^{-(T^2-1/2)} \quad (3.14)$$

From 3.12 and 3.13,

$$\frac{\mu_{2,(shift=T)}}{\mu_{2,MC}} \sim \frac{e^{-\frac{T^2}{2}}}{2} \xrightarrow{T\longrightarrow\infty} 0 \qquad (3.15)$$

Again,

$$\frac{\mu_{2,(shift=T)}}{\mu_{2,(scale=T)}} \sim \frac{e^{-1/2}}{T} \xrightarrow{T\longrightarrow\infty} 0 \qquad (3.16)$$

From 3.15 and 3.16, we infer that asymptotically shifting performs infinitely better than both the scaling and the crude methods.

General case :

Again, we directly evaluate the ratio of the second moments as follows :

$$\lim_{T\longrightarrow\infty} \frac{\mu_{2,(shift=T)}}{\mu_{2,MC}} = \lim_{T\longrightarrow\infty} \frac{\int_T^\infty \frac{f^2(x)}{f(x-T)}\,dx}{\int_T^\infty f(x)\,dx} = \lim_{T\longrightarrow\infty} \frac{f(T)}{f(0)} = 0$$

$$\lim_{T\longrightarrow\infty} \frac{\mu_{2,(shift=T)}}{\mu_{2,(scale=T)}} = \lim_{T\longrightarrow\infty} \frac{\int_T^\infty \frac{f^2(x)}{f(x-T)}\,dx}{\int_T^\infty \frac{f^2(x)}{\frac{1}{T} f(\frac{x}{T})}\,dx} = 0$$

(by L'Hospital's rule and Leibnitz theorem).

Hence, for the general case, again, shift by T performs better than both the crude method and the scaling.

Shifting and scaling :

In the previous section, for the shifted and scaled version of $p_o(x)$, we obtained the following solution

$$\theta_o = \mu_{1,opt}$$

and $$0 < \sigma_o^2 < 2 + \frac{1}{T^2-1}$$

or asymptotically $\theta_o \sim T$, $\sigma_o^2 \in (0,2)$.

Let us further investigate this biasing technique for the Gaussian family. Computing the second moment of the Importance Sampling estimator with $p_{\theta_o,\sigma}(x)$ as the biasing density, with $\theta_o \sim T$ and some arbitrary $\sigma > 0$, we obtain,

$$\mu_{2(\theta_o,\sigma)} = \int_T^\infty \left[ \frac{p_o(x)}{p_{\theta_o,\sigma}(x)} \right]^2 p_{\theta_o,\sigma}(x)\, dx$$

$$\sim \sigma \cdot \frac{1}{\sqrt{2\pi}} \frac{e^{-T^2}}{2T} \qquad (3.17)$$

From 3.17 and 3.13, we get

$$\frac{\mu_{2(\text{shift and scale})}}{\mu_{2(\text{shift})}} \sim \sigma.$$

Hence for any value of $\sigma << 1$, the shifted and scaled version performs better than the shifted version. This is an intuitively satisfying result.

In these two sections, we have managed to put the various biasing schemes under an unified framework and improved upon the existing results for some parametric classes, especially for the Gaussian case.

## 3.3 ASYMPTOTIC OPTIMALITY OF SHIFT THROUGH THRESHOLD :

Let us form a Huber neighbourhood (2.8) as follows : Put $\varepsilon_1 = 0$ and $p_1 = p_{opt}$ and $p_o = \{q | q = p(x-\theta), \theta \in (0,\infty)\}$. Now, suppose the least favourable pair $(q_o, p_{opt})$ exists. Then by Kobayashi's theorem,

$$q_o = \arg \min_{q_o' \in \mathbb{P}_o} d_{IS}(q_o', p_{opt})$$

But, counter examples in section 2.3 have shown the non-existence of this pair, at least for the double exponentialand the exponential classes, for finite T. Here, we prove that for $T \longrightarrow \infty$, this pair does exist for any shift class, under suitableconditions. To prove this, we use Huber's lemma restated as theorem 2 in section 2.2. We rewrite this here for convenience :

If a pair $(q_o, q_1)$ satisfies,

$$Q_o' \left[ \frac{q_1}{q_o} < t \right] \geq Q_o \left[ \frac{q_1}{q_o} < t \right] \geq Q_1 \left[ \frac{q_1}{q_o} < t \right] \geq Q_1' \left[ \frac{q_1}{q_o} < t \right], \tag{3.18}$$

for any real t, then it is a least favourable pair.

**Theorem 4 :**

If $p_1 = \{p_{opt}\}$

and $p_o = \{q \mid q = p(x-\theta),\ \theta \in (0,\infty)\}$

where

$$p_{opt}(x) = \frac{I_{[T,\infty)}(x)\ p(x)}{k}, \quad k = \int_T^\infty p(x)\, dx,$$

then the pair $(p(x-T), p_{opt}(x))$ is least favourable asymptotically, as $T \longrightarrow \infty$ under the following conditions :

i) $p(.)$ is differentiable.

ii) $\frac{p(x)}{p(x-\theta)}$ is monotone decreasing wrt x for $\theta > 0$ and is bounded above.

**Proof :**

We see that for this class, the right most two inequalities in (3.18) are trivially satisifed. We now assume that for some $T_o = T_o(T)$, $p(x - T_o) = q_o$. Then, let us minimise

$$I = \int p(x - f(T))\, dx \quad \text{wrt } f(T), \text{ for all } T \text{ and } t.$$

$$\left[ \frac{p_{opt}(x)}{p(x - T_o)} < t \right]$$

Now, given a t, there exists an $x_o = \varphi(T_o,t)$ such that on $\Lambda = \{x : x \geq \varphi(T_o,t)\}$,

$$\frac{p_{opt}(x)}{p(x - T_o)} < t \quad \text{(by condition ii)}$$

Then,

$$I = \int_{-\infty}^{T} p(x - f(T))\, dx + \int_{\varphi(T_o,t)}^{\infty} p(x-f(T)\, dx$$

is to be minimised wrt f(T) for all T. Now,

$$I' = p(x-f(T))[1/f'(T) - 1] + p[\varphi-f(T)][1-\varphi'(T_o,t)] = 0 \quad \text{for all } T.$$

$$\Longrightarrow \quad f(T) = T + C_1 \quad \text{and } \varphi(T_o,t) = T + C_2$$

Thus, asymptotically $\varphi(T_o,t)$ is independent of t and Lim $f(T)/T$ $T \to \infty$ $\to$ 1.

QED.

The double-exponential family does not satify the above condition (no. ii). Hence, we give a separate proof for this here.

**Proposition :**

Theorem 4 holds for the double exponentialfamily, without condition (ii).

**Proof :**

$$\text{Here, } p_{opt}(x) = e^{T-x}, \quad x \geq 0$$

$$p(x - \theta) = \frac{1}{2} e^{-|x-\theta|}$$

First consider the case when $\theta \in (0,T]$ :

$$\left\{ x : \frac{p_{opt}(x)}{p(x-\theta)} < t \right\} = \begin{cases} (-\infty, T], & t < 2e^{T-\theta} \\ (-\infty, \infty), & t \geq 2e^{T-\theta} \end{cases}$$

For $t \geq 2e^{T-\theta}$, the inequality holds trivially. For $t < 2e^{T-\theta}$, we minimise

$$I = \int_{-\infty}^{T} \frac{1}{2} \cdot e^{x-f(T)}\, dx \quad \text{wrt } f(T), \forall\, T.$$

This gives,

$$I' = \frac{1}{2}\,[1 - f'(T)]\, e^{T-f(T)} = 0, \forall\, T.$$

$$\Longrightarrow \quad f(T) = T + C \quad \text{and} \quad \lim_{T \to \infty} \frac{f(T)}{T} = 1$$

For the next case, i.e, $\theta \in (T,\infty)$,

$$\frac{p_{opt}(x)}{p(x-\theta)} = \begin{cases} 0, & x < T \\ 2e^{T+\theta-2x}, & \theta \geq x \geq T \\ 2e^{T-\theta}, & x \geq \theta \end{cases}$$

Hence, we minimise

$$I = \int_{-\infty}^{T} \frac{1}{2}\, e^{x-f(T)}\, dx + \int_{\frac{T+T_o}{2} - \frac{1}{2}\ln\frac{t}{2}}^{f(T)} \frac{1}{2}\, e^{x-f(T)}\, dx + \int_{f(T)}^{\infty} \frac{1}{2}\, e^{f(T)-x}\, dx$$

for all T. Here, $p(x-T_o(T))$ is used to denote the member that is least favourable with $p_{opt}(x)$. Then

$$I' = \frac{1}{2}\left[\frac{1}{f'(T)} - 1\right] e^{T-f(T)} + \frac{1}{2}\left[\frac{\frac{1}{f'(T)} + \frac{T_o'}{f'(T)}}{2} - 1\right] \cdot e^{\left(\frac{T+T_o}{2}\right) - \frac{1}{2}\ln\frac{t}{2} - f(T)} = 0$$

$$\therefore \quad f'(T) = 1, \quad T_o'(T) = 1, \quad \forall\ T.$$

$$\text{i.e.} \quad \lim_{T\to\infty} \frac{f(T)}{T} = \lim_{T\to\infty} \frac{T_o(T)}{T} = 1$$

QED.

The asymptotic optimality of shift through threshold has previously been established only throughad hoc approaches and brute-force simulation. See orsak and Aazhang (1989), Wessel et al. (1988), Beaulieu (1990), for example. Here, using Huber's lemma and Kobayashi's theorem, we have proved it rigorously.

# CHAPTER - IV

# SEQUENTIAL TESTS

Siegmund (1976) has applied importance sampling in the Monte Carlo estimation of error probabilities in sequential tests. In this chapter, we study a very general class of sequential tests called Sequential Minimum Probability Ratio Tests introduced by Hall (1980) and guided by Siegmund's results, apply importance sampling to evaluate error probabilities in this case. First, we briefly review Hall's paper (1980).

## 4.1 HALL'S SEQUENTIAL MINIMUM PROBABILITY RATIO TESTS (MPRTs):

Let $f_{in}$ $(i = 0,1,2)$ be distinct alternative joint densities of data $X_{(n)} = (X_1, \ldots, X_n)$ defined consistently for $n = 1,2,\ldots$ wrt a common dominating measure $\mu_n$. Call $f_{on}$ the "intermediary" hypothesis chosen for the test of $f_{1n}$ vs. $f_{2n}$. Let $d_i$ $(i = 1,2)$ be the decision that $(f_{in})$ is correct. Choose $\lambda_i$ $(i = 1,2) > 0$ such that $\sum \lambda_i = 1$. Define

$$l_n \overset{\Delta}{=} \min_{i=1,2} \left( \lambda_i \frac{f_{in}}{f_{on}} \right).$$

Hall's MPRT is a random vector (N,D) defined as follows :

$$\text{Stopping rule } N: \quad N = \inf \{n : l_n \leq \alpha \ (< 1)\}$$

$$\text{Decision rule } D: \quad \lambda_1 f_{1n} \underset{d_2}{\overset{d_1}{\gtrless}} \lambda_2 f_{2N} \qquad (4.1)$$

In the above, (N,D) describes a very wide class of sequential tests including Wald's SPRT, Lorden's 2-SPRT, Anderson's tests, sequential robust tests etc. as special cases. In particular when $H_o$, $H_1$ and $H_2$ are iid, we have Lorden's tests. The main advantage of the MPRT is that it allows control over the weighted average of the two error probabilities. This can be proved as follows.

Let $S_{in}$ be the subset of the range of $X_{(n)}$ in which N = n and $d_i$ is chosen. In $S_{2n}$, $\lambda_1 f_{1n} \leq \lambda_2 f_{2n}$ so that $\lambda_1 f_{1n} = l_n f_{on} \leq \alpha f_{on}$. Hence,

$$\lambda_1 P_1(d_2) = \sum_n \int_{S_{2n}} \lambda_1 f_{1n} \leq \sum_n \int_{S_{2n}} \alpha f_{on} = \alpha P_o(d_2) \qquad (4.2A)$$

and similarly,

$$\lambda_2 P_2(d_1) \leq \alpha P_o(d_1)$$

Adding the two,

$$\lambda_1 P_1 (d_2) + \lambda_2 P_2 (d_1) \leq \alpha \qquad (4.2B)$$

SPRT as a Special Case :

Let $f_{on} = \mu_1 f_{1n} + \mu_2 f_{2n}$, $\mu_i > 0$ and $\sum \mu_i = 1$. Let SPRT (B,A) be the SPRT of $H_1$ vs $H_2$, deciding in favour of $H_1(H_2)$ when $l_n \leq B$ $(\geq A)$, $0 < B < A$. W log, let $\lambda_1 = \lambda$ and $\lambda_2 = 1-\lambda$. Then, for every $\lambda \in (\frac{B}{1+B}, \frac{A}{1+A})$, there exists $(\alpha, \lambda, \mu)$ for which

$$\text{MPRT } (\alpha, \lambda, \mu) \stackrel{=}{=} \text{SPRT } (B,A).$$

Specifically,

$$\mu_1 = \frac{a(\lambda)}{a(\lambda) + b(\lambda)} \qquad (4.3a)$$

where

$$a(\lambda) = \frac{B(1+A)}{\left(\frac{A}{1+A} - \lambda\right)} \quad \text{and} \quad b(\lambda) = \frac{(1+B)}{\left(\lambda - \frac{B}{1+B}\right)} \qquad (4.3b)$$

$$\mu_2 = 1 - \mu_1 \quad \text{and} \quad \alpha = \frac{\lambda}{(A + (1-A)\mu_1)} \qquad (4.4)$$

For proof of the above and further discussions, see Hall (1980).

## 4.2 IMPORTANCE SAMPLING FOR MPRTs:

We first review Siegmund's (1976) results on importance sampling for sequential tests.

### 4.2.1 Siegmund's results for SPRTs :

Let $X_k$ be iid with a common distribution P such that

$$-\infty < EX_k < 0 \quad \text{and} \quad P\{X_k > 0\} > 0 \tag{4.5}$$

Let $S_n = \sum_1^n X_k$ and for $a \leq 0 < b$, let

$$T = \inf \{n : S_n \notin (a,b)\}.$$

To estimate $\alpha = P\{S_T \geq b\}$ using importance sampling through the choice of a biasing density from the class

$$P_\theta \{x_k \in dx\} = \exp(\theta x - \Psi(\theta))\, dH(x) \tag{4.6}$$

where $\Psi$ is normalised so that $\Psi(0) = \Psi'(0) = 0$, Siegmund gives the following result :

Let u be that value of $\theta$ such that $P = P_u$. By 4.5, $u < 0$. By the convexity of $\Psi$, there exists a $\omega > 0$ for which $\Psi(\omega) = \Psi(u)$. Then, as $b \longrightarrow \infty$, for all $\theta \neq \omega$, $\mu_2(\omega)/\mu_2(\theta)$ converges to 0 at an exponential rate where $\mu_2(\theta)$ is the second

moment of the important sampling estimate with $P_\theta$ as the biasing density.

In particular, for the case of a simple hypothesis $H_o$ against a simple alternative $H_1$, if $P_u$ is the measure corresponding to $H_o$ then, $P_\omega$ corresponds to $H_1$. See Siegmund (1980) pp 14-19. Again, when $x_k$'s are loglikelihood ratios, we have Wald's SPRT.

### 4.2.2 Biasing density for error estimation in MPRT :

Going back to the notation of 4.1, the probability of false alarm in the case of MPRT can be written (4.2A) as

$$\alpha_1 = P_1(d_2)$$

$$= \int_{S_{2N}} f_{1N}(x_1, \ldots\ldots, x_n)\, dv^{(N)} \qquad (4.7)$$

where $dv^{(N)} = dx_1 \ldots\ldots dx_N$ and

$$S_{2N}(x_1,\ldots\ldots,x_N) = \{(x_1,\ldots\ldots,x_N) : l_N < \alpha$$
$$\text{and } \lambda_1 f_{1N}(x_1,\ldots\ldots,x_N) \leq \lambda_2 f_{2N}(x_1,\ldots\ldots,x_N)\}$$

Recall $N = \inf\{n : l_n < \alpha\}$.

$$\therefore \quad \alpha_1 = \int f_{1N}(x_1,\ldots,x_N)\, dv^{(N)} \qquad (4.8)$$

$$\left\{ \frac{f_{0N}(x_1,\ldots,x_N)}{f_{1N}(x_1,\ldots,x_N)} \geq \frac{\lambda_1}{\alpha} \; ; \; \frac{f_2(x_1,\ldots,x_N)}{f_{1N}(x_1,\ldots,x_N)} \geq \frac{\lambda_1}{\lambda} \right\}$$

The unbiased crude Monte Carlo estimate for $\alpha_1$ when $f_{1N}$ is sampled can be written as

$$\hat{\alpha}_1 = \frac{1}{N} \sum_{k=1}^{N} I\left\{ \log\left( \frac{f_{0N_k}(\overline{X}_k)}{f_{1N_k}(\overline{X}_k)} \right) \geq k_1 \; ; \; \log\left( \frac{f_{2N_k}(\overline{X}_k)}{f_{1N_k}(\overline{X}_k)} \right) \geq k_2 \right\}$$

where $I(B)$ is the Indicator on the Borel set $B \in \mathbb{B}_k$, the Borel field over $\mathbb{R}^k$,

$$k_1 \overset{\Delta}{=} \log\left( \frac{\lambda_1}{\alpha} \right) \text{ and}$$

$$k_2 \overset{\Delta}{=} \log\left( \frac{\lambda_1}{\alpha} \right), \quad \text{and} \quad \overline{X}_k \overset{\Delta}{=} X_{(k)}$$

Let us consider the following class of densities for $X_{(n)}$, defined analogous to the $P_\theta$ class (4.6) :

$$f^*_{t_1,t_2}(x_1,\ldots,x_n) = \exp[t_1 Z_{1n}(x_1,\ldots,x_n) + t_2 Z_{2n}(x_1,\ldots,x_n) - \Psi_n(t_1,t_2)]\, f_{on}(x_1,\ldots,x_n) \qquad (4.9)$$

where

$$Z_{1n}(.) = \log\left[\frac{f_{on}(.)}{f_{1n}(.)}\right] \text{ and } Z_{2n}(.) = \log\left[\frac{f_{2n}(.)}{f_{1n}(.)}\right]$$

and

$$e^{\Psi_n(t_1,t_2)} = \varphi_n(t_1,t_2) = \int_{\mathbb{R}^n} \exp\,[t_1\, Z_{1n}(x_1,\ldots,x_n) + t_2\, Z_2(x_1,\ldots,x_n)]\, f_{on}(x_1,\ldots,x_n)\, dv^{(n)}.$$

It is easy to see that this class includes $f_{on}$, $f_{1n}$ and $f_{2n}$. Specifically,

$$\Psi_n(0,0) = \Psi_n(-1,\,0) = \Psi_n(-1,\,1) = 0 \quad \text{and}$$

$$f^*_{0,0}(x_1,..,x_n) = f_{on}(x_1,..,x_n),\ f^*_{-1,0}(x_1,..x_n)$$

$$= f_{1n}(x_1,..,x_n) \text{ and } f^*_{-1,1}(x_1,..,x_n) = f_{2n}(x_1,\ldots,x_n).$$

We also observe that the actual choice of $f_{in}$ $(i = 0,1,2)$ can be arbitrary.

**Theorem 5 :**

Let $\mu_2(t_1',\ t_2')$ be the second moment of the importance sampling estimate of $\alpha_1$ with $f^*_{t_1',t_2'}$ as the biasing density. Then, $\mu_2(-1,\ 1)/\mu_2(-1,0)$ converges to zero exponentially as $k_2 \longrightarrow \infty$ .

**Remarks :**

i) The choice of $f_{2n}$ as the biasing density ($X_{(n)}$ is sampled from $f_{2n}$ instead of $f_{1n}$) and the consequent use of the estimate

$$\hat{\alpha}_1^* = \frac{1}{n} \sum_{k=1}^{n} \frac{f_{1N_k}}{f_{2N_k}} \quad I\left\{ \log\left( \frac{f_{0N_k}}{f_{1N_k}} \right) \geq k_1 ; \log\left( \frac{f_{2N_k}}{f_{1N_k}} \right) \geq k_2 \right\}$$

where $f_{iN_k} \overset{\Delta}{=} f_{iN_k}(\bar{X}_k)$, $i = 0, 1, 2,$

gives an unbiased estimate of $\alpha_1$ at an exponentially faster rate as compared to the use of $\hat{\alpha}_1$.

ii) The justification for choosing $k_2$ as the asymptotic parameter and not $k_1$, is as follows : From 4.1, we see that $k_1 = \log(\lambda_1/\alpha)$ is the termination threshold and $k_2 = \log(\lambda_1/\lambda_2)$ is the decision threshold. It is our aim here, as it has been in all asymptotic works on importance sampling, to evaluate the order of reduction in the sample-size asymptotically as the quantity to be evaluated goes to zero. For the MPRT, the probability of false alarm $\alpha_1$ decreases as the margin of decision $k_2$ increases. Therefore, whereas allowing $k_2 \longrightarrow \infty$ results in $\alpha_1 \longrightarrow 0$, $k_1 \longrightarrow \infty$ will only affect the value of E(N).

Formally, let $\sigma(l_n)$ be the $\sigma$- field generated by the random variable

$$l_n = \min (X,Y), \text{ where } X = \frac{f_{on}}{f_{1n}} \text{ and } Y = \frac{f_{2n}}{f_{1n}} .$$

Then the decision event, which is an event of the form $\{X \leq Y\}$ or $\{X > Y\}$ is not measurable wrt $\sigma(l_n)$.

**Proof of Theorem 5 :**

This follows by the direct computation of the moments : If $f^{*}_{(t_1',t_2')}$ is the biasing density, then,

$$\alpha_1 = \int_{\{Z_{1N}(x_1,..,x_N) \geq k_1 \; ; \; Z_{2N}(x_1,.....,x_N) \geq k_2\}} \exp [t_1 Z_{1N}(x_1,...,x_N) + t_2 Z_{2N}(x_1,..,x_N) - \Psi_N(t_1,t_2)] \, d F_{on}(x_1,..,x_n).$$

$$= \int_{\{Z_{1N}(x_1,..,x_N) \geq k_1 \; ; \; Z_{2N}(x_1,..,x_N) \geq k_2\}} \exp [(t_1-t_1') Z_{1N}(x_1,...,x_N) + (t_2-t_2') Z_{2N}(x_1,..,x_N) - (\Psi_N(t_1,t_2) - \Psi_N(t_1',t_2'))] \, dF_{t_1,t_2}(x_1,....,x_N) \tag{4.10}$$

The second moment of the estimate of $\alpha_1$ as expressed in (4.10) above is :

$$\mu_2(t_1',t_2') = \int \exp\,[(2t_1-t_1')\,Z_{1N}(x_1,..,x_N) + (2t_2-t_2')Z_{2N}(x_1,..x_N) - \{2\Psi_N(t_1,t_2) - \Psi_N(t_1',t_2')\}]\; dF_{on}(x_1,..x_N)$$

$$\{Z_{1N}(x_1,....,x_N) \geq k_1;\; Z_{2N}(x_1,...,x_N) \geq k_2\}$$

Then the result directly follows :

$$\mu_2(-1,1) = \int \exp\,[-Z_{1N} - Z_{2N}]\; dF_{ON}$$

$$\{Z_{1N} \geq k_1\;;\; Z_{2N} \geq k_2\}$$

$$\leq \exp\,(-k_2) \int [-Z_{1N}]\; dF_{ON}$$

$$\{Z_{1N} \geq k_1\;;\; Z_{2N} \geq k_2\}$$

$$= \exp\,(-k_2)\;\mu_2\,(-1,0).$$

QED

AN EXAMPLE :

Let us consider the Gaussian case of $f_o(x) = N(o,1)$, $f_1(x) = N(-\mu,1)$ and $f_2(x) = N(+\mu,1)$ where $H_o$, $H_1$ and $H_2$ are iid. Then,

$$Z_1(x) = \frac{\mu^2}{2} + \mu x \quad \text{and} \quad Z_2(x) = 2\mu x$$

$$\Psi(t_1,t_2) = \frac{\mu^2}{2}\,[t_1 + (t_1 + 2t_2)^2$$

$\therefore \Psi(0,0) = \Psi(-1,0) = \Psi(-1,1) = 0$

$\mu_2(-1,1) \leq \exp(-k_2)\, \mu_2(-1,0)$ is also verified.

### 4.2.3 SPRT as a special case of MPRT :

When, under the conditions listed in (4.3) and (4.4), MPRT reduces to SPRT, we show that the result in Theorem 5 remains consistent.

**Lemma :** With $f_{0n} = \mu_1 f_{1n} + \mu_2 f_{2n}$ and conditions (4.3) and (4.4), when SPRT $(B,A) \equiv$ MPRT $(\alpha, \lambda, \mu)$, $A \longrightarrow \infty$ if $k_1$ and $k_2 \longrightarrow \infty$.

**Proof :** Here,

$$k_1 = \log\left[\frac{\lambda}{\alpha}\right] \text{ and } k_2 = \log\left[\frac{\lambda}{1-\lambda}\right] \qquad (4.12)$$

Given, $\alpha = \dfrac{\lambda}{[A + (1-A)\,\mu_1]}$,

substituting for $\mu_1$ from 4.3, we get,

$A = \dfrac{\lambda + B(\alpha-1)}{\alpha + B(\lambda-1)}$. Keeping B fixed,

$$A \longrightarrow \infty \quad \text{when} \quad \lambda \longrightarrow 1 \quad \text{and} \quad \alpha \longrightarrow 0. \qquad (4.13)$$

From (4.12) and (4.13), the result follows.

QED.

As $k_1 \longrightarrow \infty$ and $k_2 \longrightarrow \infty$, the MPRT termination and decision rules merge into the SPRT termination - decision rule for $f_{on} = \mu_1 f_{1n} + \mu_2 f_{2n}$ and remain consistent :

From (4.3) and (4.4) we have $\mu_1 = \mu$, $\mu_2 = 1-\mu$ where

$$\mu = \frac{B(1+A)\left[\frac{A}{1+A} - \lambda\right]}{B(1+)\left[\frac{A}{1+A} - \lambda\right] + (1+B)\left[\lambda - \frac{B}{1+B}\right]}$$

Then, as $A \longrightarrow \infty$, $\mu \longrightarrow 1$ $\therefore f_{on}$ merges with $f_{2n}$. Further, for the choice $f_{on} = \mu f_{1n} + (1-\mu) f_{2n}$, the termination rule reduces to

$$\left(\frac{f_{2n}}{f_{1n}}\right) \geq \frac{1}{(1-\mu)}\left[\frac{1}{B}\left(\frac{\lambda}{1-\lambda}\right) - \mu\right]$$

and the decsion rule to

$$\left(\frac{f_{2n}}{f_{1n}}\right) \geq \left(\frac{\lambda}{1-\lambda}\right),$$

which for $\lambda \longrightarrow 1$ imply that the upper threshold $A \longrightarrow \infty$.

# CHAPTER - V

## CONCLUSION

For some parametric families, we have shown that the solution to the equation

$$P^{*} = \arg \min_{P \in \mathbb{P}} d(P_{opt}, P)$$

is easier to obtain with the Kullback-Leibler distance than with the importance sampling distance as a choice for d. Orsak and Aazhang (1990) have considered the mixture neighbourhood for $\mathbb{P}$. but the least favourable density is dependant on $P_{opt}$. A suitable implementable modification of this density is desirable. One can further consider other neighbourhoods like the total variational neighbourhood, which are topological neighbourhoods and are therefore much fuller and may yield better solutions. But, here again, the solutions should be modified and made independant of $P_{opt}$ since it depends on the (unknown) quantity to be estimated.

We have established the asymptotic optimality of a member of the location family thatis shifted by the threshold. Similar result for the location-scale family is desirable since it performs better than the location family.

In case of MPRTs, we have only shown that the use of the density of the alternative hypothesis for biasing gives a large reduction in the sample-size. The question of optimality remains open. Further, one still needs to consider the multiple decision case as discussed in Hall (1980).

# REFERENCES


1. S.M. Ali and D. Silvey (1966), 'A General Class of Coefficients of Divergence of One Distribution from Another', J. Royal Stat. Soc., Vol. 28, pp 131-142.
2. N.C. Beaulieu (1990), 'A Composite Importance Sampling Technique for Digital Communication System Simulation', IEEE Trans. Commn., Vol. 38, No. 4, pp 393-396.
3. M. Cottrell, J-C. Fort and G. Malgouyres (1983), ' Large Deviations and Rare Events in the Study of Stochastic Algorithms', IEEE Trans. Auto. Control, Vol. AC-28, No. 9, pp 907-920.
4. P.M. Hahn and N.C. Jeruchim (1987), 'Developments in the Theory and Applications of Importance Sampling', IEEE Trans. Commn., Vol. COM-35, No. 7, pp 706-713.
5. W.J. Hall (1980), 'Sequential Minimum Probability Ratio Tests', in ASYMPTOTIC THEORY OF STATISTICAL TESTS AND ESTIMATION, Ed. I.M. Chakraborty, Acad. Press, pp 325-350.
6. J.M.. Hammersley and D.C. Handscomb (1964), MONTE CARLO METHODS, Methuen, NY.
7. P.J. Huber (1965), 'A Robust Version of the Probability Ratio Tests', Ann. Math. Stat., Vol. 36, pp 1753-1758.
8. H. Kahn and A.W. Marshall (1953), 'Methods of Reducing Sample Size in Monte Carlo Computations', J. Op. Res. Soc. Am., Vol. 1, pp 263-278.

9. M.H. Kalos and P.A. Whitlock (1986), MONTECARLO METHODS, Vol. I, John Wiley, NY.

10. H. Kobayashi (1967), 'Distance Measures and Related Criteria', Proc. 5th Allerton Conference on Circuit and System Theory, OCT. 67, pp 491-500.

11. G.W. Lank (1983), 'Theoretical Aspects of Importance Sampling Applied to False Alarms', IEEE Trans. Inform. Th., Vol. IT-29, No. 1, pp 73-82.

12. D. Lu and K. Yao (1988a), 'Improved Importance Sampling Technique for Efficient Simulation of Digital Communication Systems', IEEE J. Selected Areas. Commn., Vol. SAC-6, pp 67-75.

13. D. Lu and K. Yao (1988b), 'On Some New Importance Sampling Results for Simulation of Non-linear Digital Communication Systems', Proceedings of the 1988 Conf. on Inform. Sci. Systems, Princeton University, Princeton.

14. G. Orsak and B. Aazhang (1988), 'A Comparison of Two Importance Sampling Methods for the Analysis of Detection Systems', Proceedings of the 1988 Conf. on Inform. Sci. Systems, Princeton University, Princeton.

15. G. Orsak and B.. Aazhang (1990), 'Constrained Solutions in Importance Sampling Via Robust Statistics', unpublished Report.

16. G. Orsak and B. Aazhang (1989), 'On the Theory of Importance Sampling Applied to the Analysis of Detection Systems', IEEE Trans. Commn., Vol. 37, No. 4, pp 332-339.

17. J.S. Sadowsky and J.A. Bucklew (1990), 'On Large Deviations Theory and Asymptotically Efficient Monte Carlo Estimation', IEEE Trans. Inform. Theory, Vol. 36, No. 3, pp 579-588.

18. K.S. Shanmugam and P. Balaban (1980), 'A Modified Monte Carlo Simulation Technique for the Evaluation of Error Rate in Digital Communication Systems', IEEE Trans. Commn., Vol. COM-28, No. 11, pp 1916-1924.

19. D. Siegmund (1976), 'Importance Sampling in the Monte Carlo Study of Sequential Tests', Ann. Stat., Vol. 4, No. 4, pp 673-684.

20. D. Siegmund (1985), SEQUENTIAL ANALYSIS, Springer-Verlag.

21. A Wald (1947), SEQUENTIAL ANALYSIS, Dover.

22. E.S. Wentzel (1988), OPERATIONS RESEARCH, Mir Publishers, Moscow.

23. A.E. Wessel, E.B. Hall and G.L. Wise (1988), 'Some Comments on Importance Sampling', Proceedings of the 1988 Conf. on Inform. Sci. Systems, Princeton University, Princeton.

24. W. Feller, AN INTRODUCTION TO PROBABILITY THEORY, Vol. I, Academic Press, 1966.

# APPNDIX - A

Here, we outline the basic properties of Ali-Silvey distances. For further details, see Ali and Silvey (1966).

Let $(\Omega, \mathbb{F})$ be a measurable space and $P_1$, $P_2$ two probability measures on it. Let $\phi$, the generalised Radon-Nikodym derivative of $P_2$ with respect to $P_1$ exist. If C is a Borel-measurable continuous convex function of a real variable and f be a Borel-measurable increasing real-valued function of a real variable, the coefficient

$$d(P_1,P_2) = f\,[E^{*}\,\{C(\phi)\}]$$

$$= f\left[\int_{\phi<\infty} C(\phi)\, dP_1 + P_2(N) \lim_{\phi\longrightarrow\infty} \frac{C(\phi)}{\phi}\right],$$

where N is a $P_1$-null set, is called the Ali-Silvey distance measure of divergence of $P_2$ from $P_1$ in the sense that it enjoys the following 4 properties :

i) $d(P_1,P_2)$ is defined for all pairs of measures $P_1$ and $P_2$ on the same sample space.

ii) If $y = t(x)$ is a measurable transformation from $(\Omega,\mathbb{F})$ onto a measure space $(\mathcal{Y},\mathcal{G})$, then,

$$d(P_1,P_2) \geq d(P_1t^{-1}, P_2t^{-1})$$

where $P_it^{-1}$ is the induced measure on $\mathcal{Y}$.

iii) $d(P_1,P_2)$ is minimum for $P_1 = P_2$ and maximum for $P_1 \perp P_2$.

iv) Let $\theta$ be a real parameter and let $\{P_\theta : \theta \in (a,b)\}$ be a family of mutually absolutely continuous distributions on the real line such that the family of densities $p_\theta(x)$ wrt a fixed measure $\mu$ has a monotone libelihood ratio in x. Then if $a < \theta_1 < \theta_2 < \theta_3 < b$, we have

$$d(P_{\theta_1}, P_{\theta_2}) \leq d(P_{\theta_1}, P_{\theta_3}).$$

A110742

EE-1991-M-GOP-IMP

1. Model sampling
2. Monte Carlo method